Student project – Regression analysis
Sang Jin Park

1. Data description

The data set that the student project covers is seismic events bigger than 4MB occurred near Fiji island since 1964. This project consists of 5 variables and 1000 observations and the explanation for each variable is provided in Table 1. The data set is embedded in statistical software R named quake and I used R for the project. In near Fiji area, there are frequent earthquakes. The attached photo is satellite photograph of Fiji Island in Figure 1 so that we can connect the earthquake data and geographic area. The right picture in Figure 1 plotted the earthquake frequencies based on longitude and latitude according to the data. This project will set models and estimate parameters to measure the effect of longitude, latitude and depth to Richter Magnitude according to the data set.

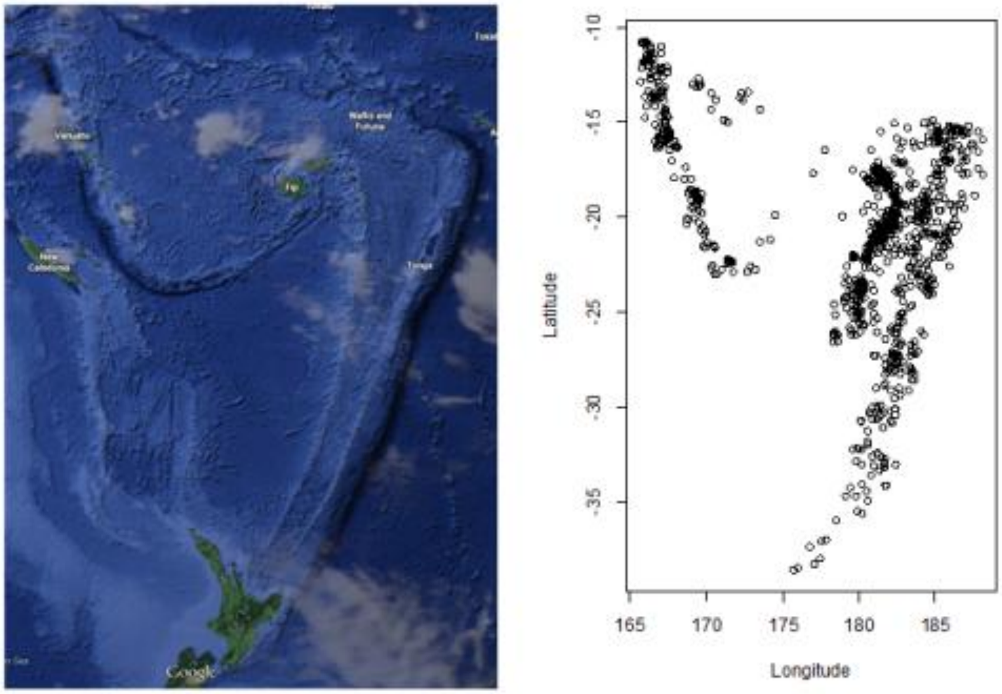


Figure 1. Satellite photograph and observations

[,1]	lat	numeric	Latitude of event
[,2]	long	numeric	Longitude
[,3]	depth	numeric	Depth (km)
[,4]	mag	numeric	Richter Magnitude
[,5]	stations	numeric	Number of stations reporting

Table 1. Description of variables

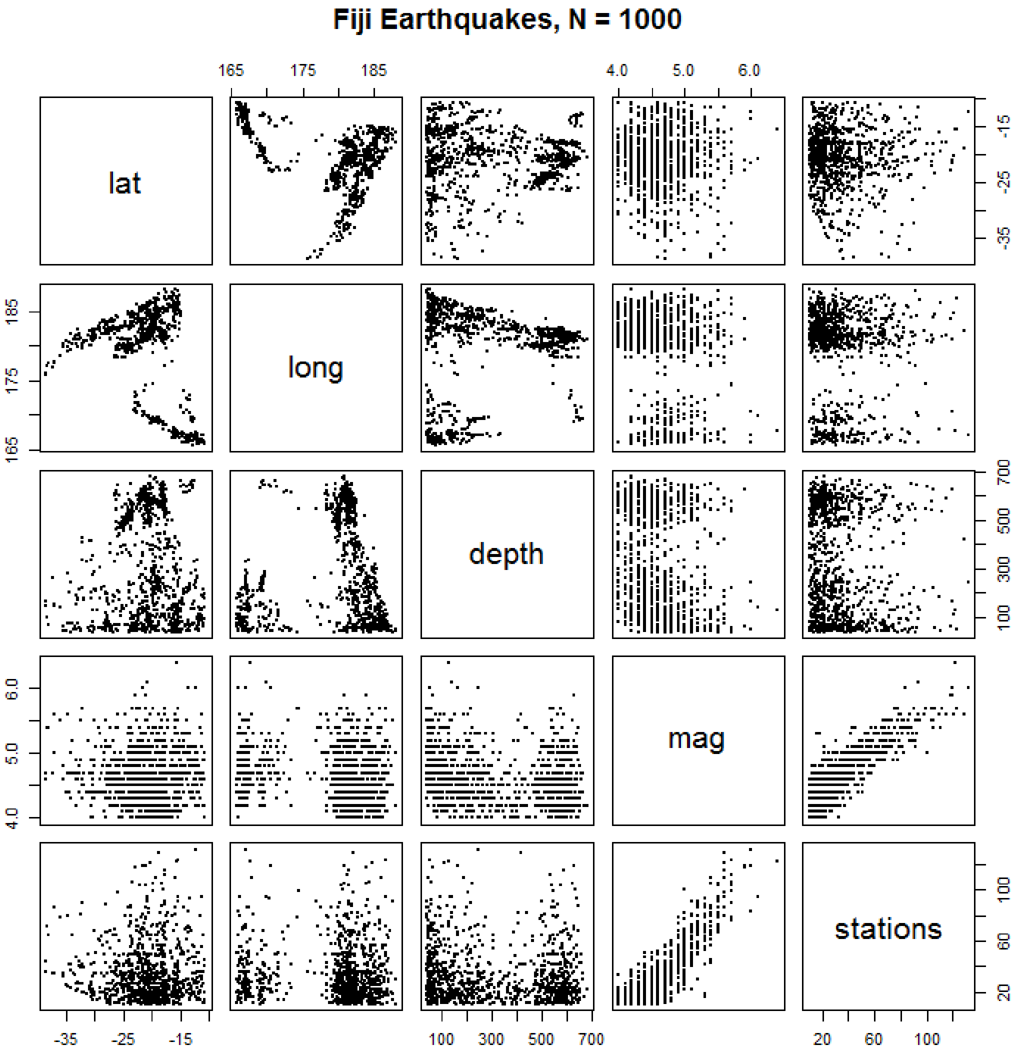


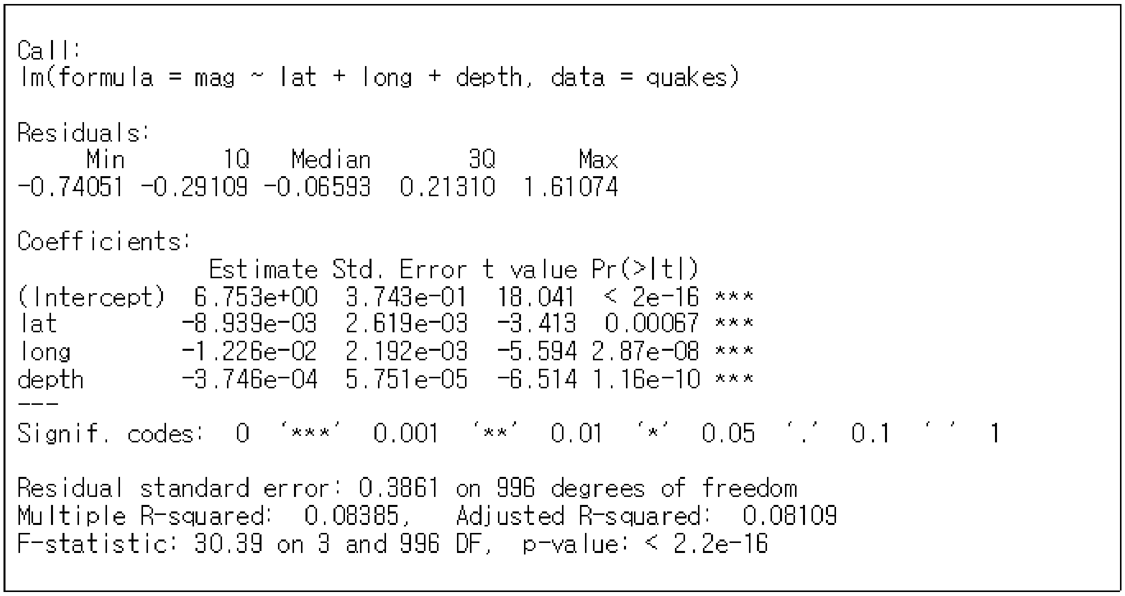
Figure 2. Scatter plot matrix

2. Model construction

Before setting regression model and perform the analysis, I drew scatter plots for each variable. The scatter plots are provided in Figure 2; it shows stations variable and mag variable show strong linearity. Stations variable means the number of stations that report earthquakes. The bigger Richter Magnitude, the more stations can detect the earthquake and this makes sense. Although other variables affect the mag variable, other variables are not as influential as mag variable. However, as stations variable are observed after the earthquakes occurred, we are going to establish regression model that explains mag with longitude, latitude and depth; I will use stations variable to adjust the effect. Let’s take a look at the regression model with three variables except stations variable.

```
fit=lm(mag~lat+long+depth,data=quakes)
summary(fit)
par(mfcol=c(2,2))
plot(fit)
```

R code 1. R code for the regression model: mag~lat+long+depth



```
Call:
lm(formula = mag ~ lat + long + depth, data = quakes)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.74051 -0.29109 -0.06593  0.21310  1.61074 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.753e+00  3.743e-01  18.041  < 2e-16 ***
lat         -8.939e-03  2.619e-03  -3.413  0.00067 ***
long        -1.226e-02  2.192e-03  -5.594 2.87e-08 ***
depth       -3.746e-04  5.751e-05  -6.514 1.16e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.3861 on 996 degrees of freedom
Multiple R-squared:  0.08385,	Adjusted R-squared:  0.08109 
F-statistic: 30.39 on 3 and 996 DF,  p-value: < 2.2e-16
```

Output 1. R output of the regression model: mag~lat+long+depth

3. Regression model with simple response

In this section, I consider a regression model that models Richter Magnitude using longitude, latitude and depth variable:

$$\mathrm{mag} = \beta_0 + \beta_1 \mathrm{lat} + \beta_2 \mathrm{long} + \beta_3 \mathrm{depth} + \epsilon$$

The R codes for regression model is provided in R code 1 and the results for the codes is provided in Output 1. When we look at Output 1, all three variables and its intercept’s coefficient of p-values are less than 0.05; this means all variables are significant. The value of coefficient of determination is 0.084 which is not very high; this means the goodness of fit of model is not significant.

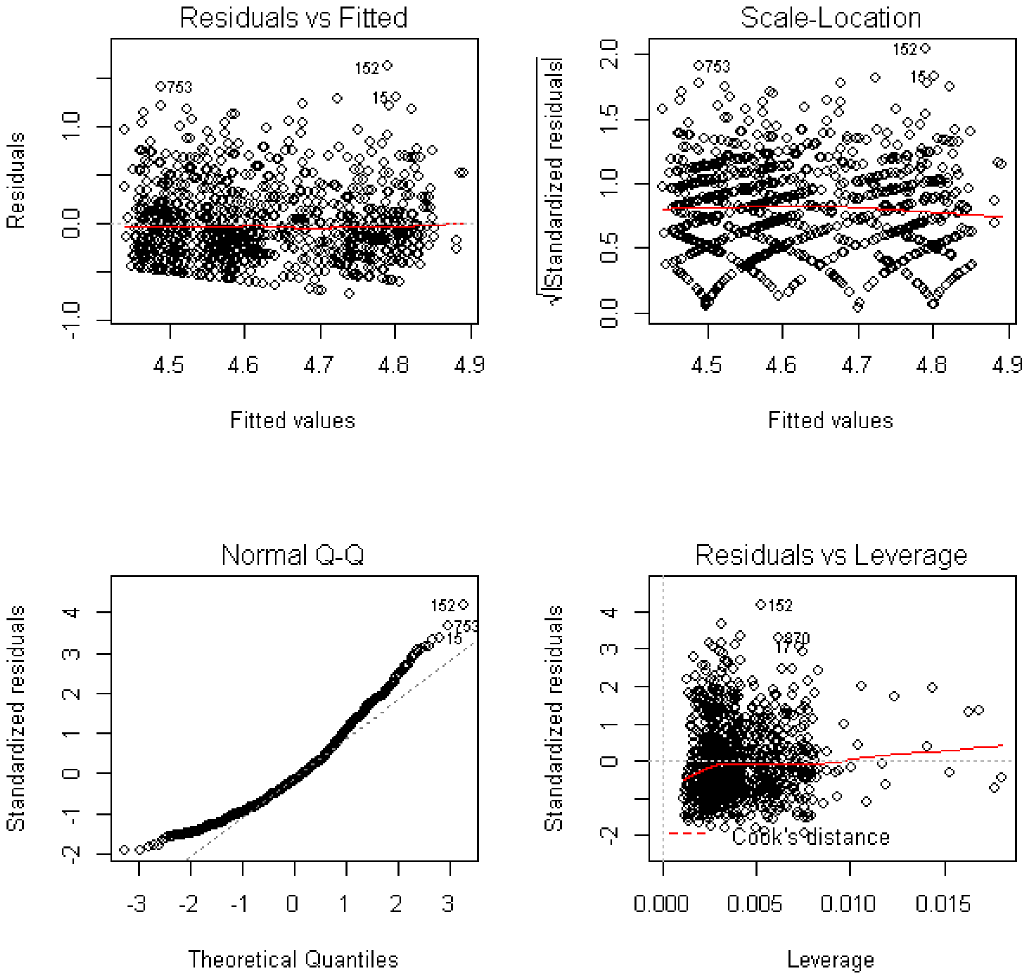


Figure 3. Diagnostic plot of the regression model: mag~lat+long+depth

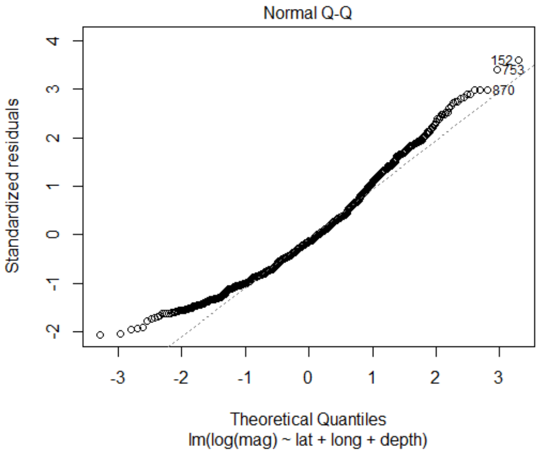


Figure 4. Q-Q plot of regression model: log(mag)~lat+long+depth

The diagnostic plot that checks whether the regression model violates the basic assumptions of regression analysis is in Figure 3. When looking at top-left panel and top-right panel, as there are no patterns of residual for fitted values and variances are maintained consistently, it satisfies the homogeneity. However, when looking at top-left panel, residuals are not symmetric, but right-skewed: this means the residual is not following the normal distribution. Bottom-left panel shows residual's Q-Q plot. Through this plot, we can observe that residuals do not follow the normal distribution and it is right-skewed. When looking at bottom-right panel, we can find out the leverage figures for residuals. Since there is no observation that Cook's distant is beyond cut-off, the problem of outlier is not occurred. As the model's residual does not follow normal distribution, we should add another variable as predictor variable to make residual follow normal distribution, or we transform variable so that residuals follow normal distribution. In next section, I do not add other variables and rather transform response variable to make residual follow normal distribution.

4. Regression model with transformations

First, for right-skewed, we can consider the log transformation. The Q-Q plot for regression model fitting results of log(mag) is provided in Figure 4. The result of log transformation is similar to the one in Q-Q plot. Therefore, I will use Box-Cox transformation which is more sophisticated than log transformation so that I can solve the problem above.

Box-transformation is defined below:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$

To estimate λ, I used box-cox function in R MASS package. The change of λ's log-likelihood is provided in Figure 5. Although λ's optimal value is about –2.75, this value is hard to be interpreted practically. So, considering λ's 95% confident interval, I set λ -3 and transform mag variable. The final model used is below:

$$\frac{\mathrm{mag}^{-3} - 1}{-3} = \beta_0 + \beta_1 \mathrm{lat} + \beta_2 \mathrm{long} + \beta_3 \mathrm{depth} + \epsilon \quad .$$

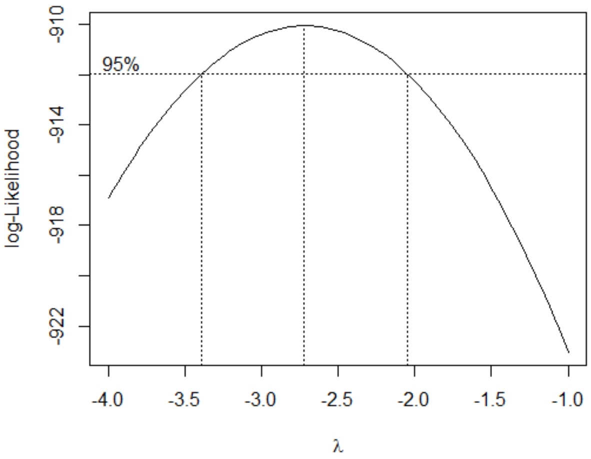


Figure 5. Estimation of lambda in the Box-Cox transformation

```
library(MASS)
boxcox(mag~lat+long+depth,data=quakes,lambda=seq(-4,-1,length=10))
```

R code 2. R code for the Box-Cox transformation

The fitted regression model and its residual Q-Q plot are provided in Figure 6.

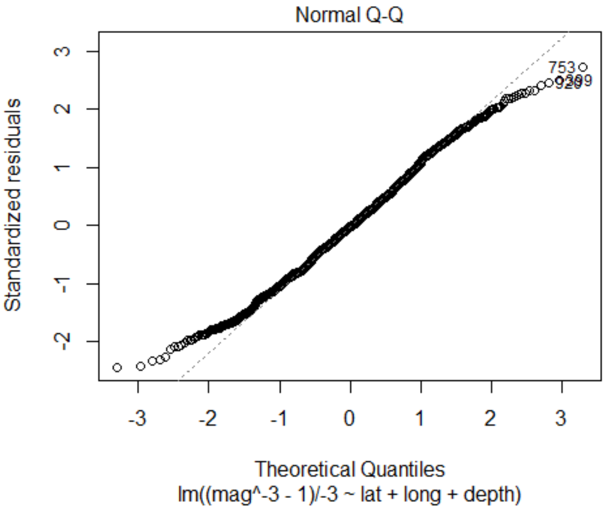


Figure 6. Q-Q plot of regression model: (mag^-3-1)/-3~lat+long+depth

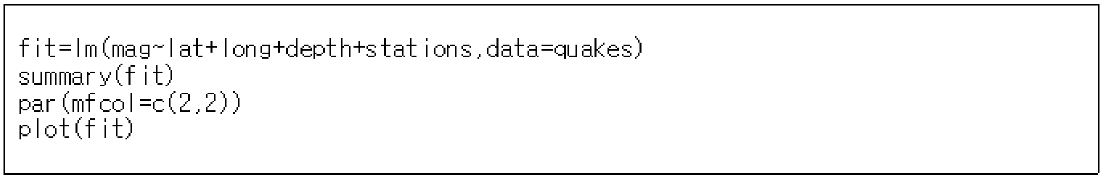
When looking at Figure 6, we can see that asymmetries are eliminated, but it has thick tails; we still cannot say it follows normal distribution. Therefore even Box-Cox transformation does not completely solve problems from residuals. We should add a predictor variable for better modeling. In next section, we are going to add station variable and provide results of adding the variable.

5. Regression model with a new predictor variable

In this section, we are going to add stations variable as an additional predictor variable and perform diagnostics. The model that we would like to consider is below:

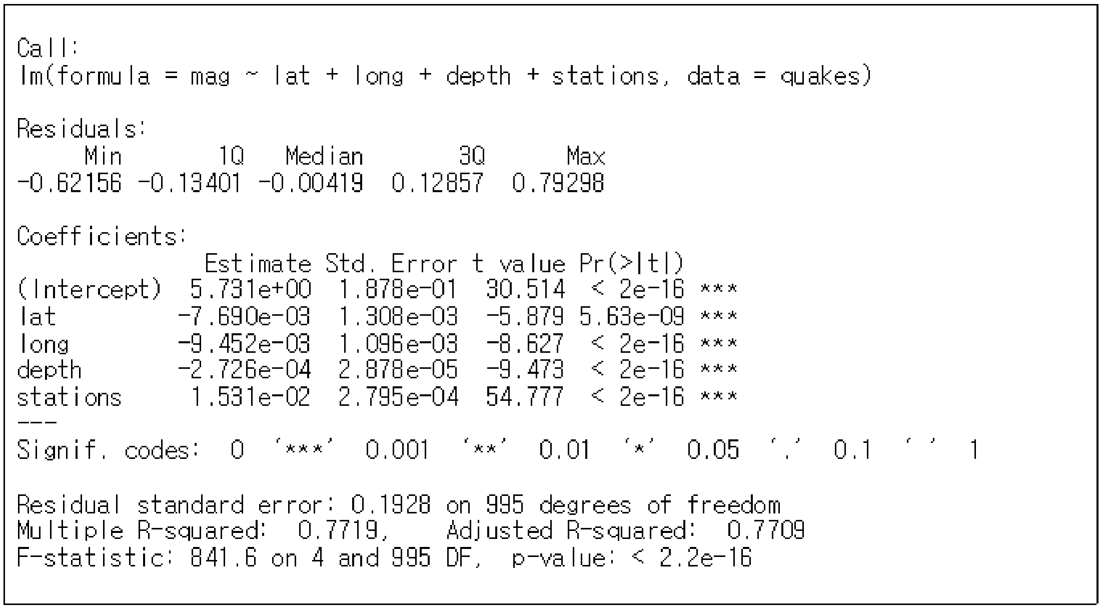
$$\text{mag} = \beta_0 + \beta_1\text{lat} + \beta_2\text{long} + \beta_3\text{depth} + \beta_4\text{stations} + \epsilon \ .$$

The R codes for fitting the model are provided in R code 3 and the results are provided in Output 2. When looking at results, we can see all variables' coefficient are significant and except stations variable, all other variables' coefficient are negative. So, the lower latitude, lower longitude, shallower depth is, the bigger Richter Magnitude is. The fact that magnitude of earthquakes is observed more significantly at shallow place matches with our common sense. Also, the relationship among latitude, longitude and mag can be estimated in Figure 1. When we look at the satellite picture in Figure1, the area that earth's plates collide which cause earthquakes exist the area whose latitude and longitude are low. The coefficients about stations are positive; the fact that number of stations that observe earthquakes means the earthquake was hugh. It matches with common sense. The model provided below tells how much of influence these variable give via β_1, β_2, and β_3 values.



```
fit=lm(mag~lat+long+depth+stations,data=quakes)
summary(fit)
par(mfcol=c(2,2))
plot(fit)
```

R code 3. R code for the regression model: mag~lat+long+depth+stations



```
Call:
lm(formula = mag ~ lat + long + depth + stations, data = quakes)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.62156 -0.13401 -0.00419  0.12857  0.79298 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.731e+00  1.878e-01  30.514  < 2e-16 ***
lat         -7.690e-03  1.308e-03  -5.879 5.63e-09 ***
long        -9.452e-03  1.096e-03  -8.627  < 2e-16 ***
depth       -2.726e-04  2.878e-05  -9.473  < 2e-16 ***
stations     1.531e-02  2.795e-04  54.777  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.1928 on 995 degrees of freedom
Multiple R-squared:  0.7719,    Adjusted R-squared:  0.7709 
F-statistic: 841.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

Output 2. R output of the regression model: mag~lat+long+depth+stations

We see the R-square is 0.77 which becomes a lot bigger from 0.08 before adding stations variable; this means the explanation of the model has been improved by adding stations variable.

The diagnostic plot for this regression model is provided in Figure 7. When looking at top-left panel and top-right panel, we can see that the skewness that existed previously is disappeared and residuals follow normal distribution symmetric. It also can be seen at bottom-left panel’s Q-Q plot; we can see that points are plotted on the line which shows residuals follow normal distribution. Therefore, we do not need to transform mag variable; it is better to use without transformation. According to bottom-right panel, there is no problem with outliers. In conclusion, the model is setup properly.

6. Conclusion

This project used earthquake data near Fiji Island to set models and estimated Richter Magnitude. As we can see in Section 4, the lower the latitude, lower longitude, shallower depth makes Richter Magnitude bigger.

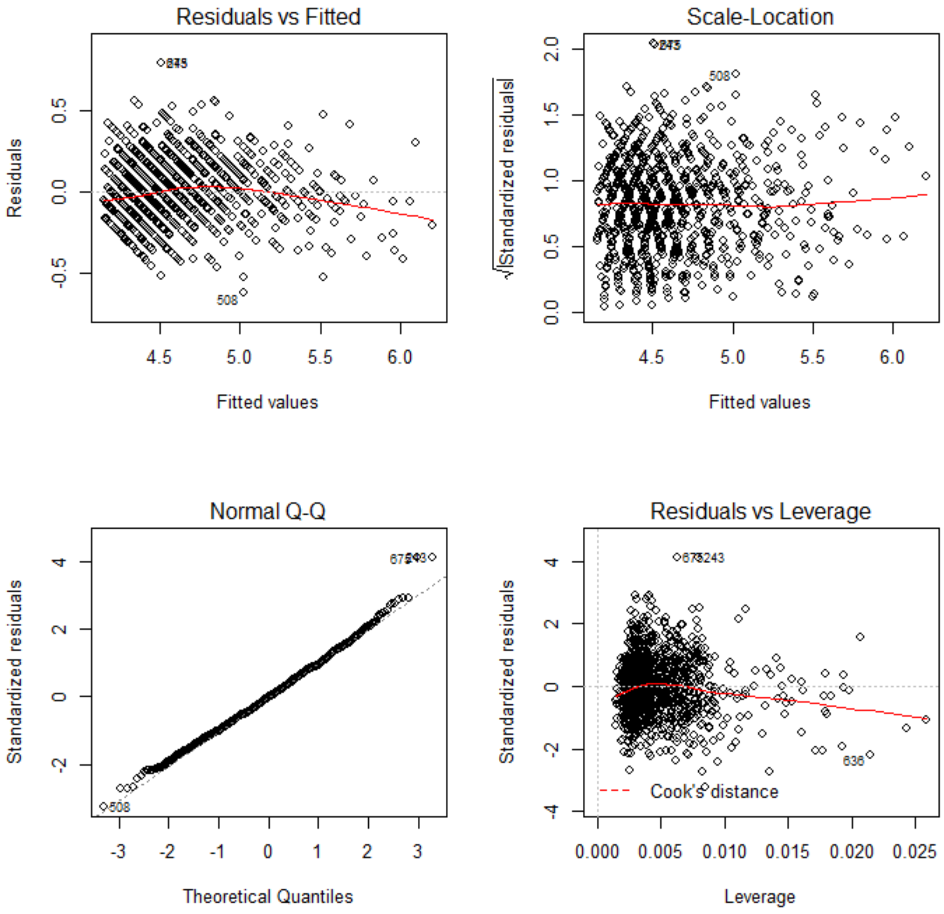


Figure 7. Diagnostic plot of the regression model: mag~lat+long+depth+stations

I made adjustments for the number of stations reporting to find out how these variables affect Richter Magnitude so that I can reduce the bias from the estimation of coefficient and maximize the model's account for observation.